SpringerBriefs in Molecular Science

Kunal Roy · Arkaprava Banerjee

q-RASAR
A Path to Predictive Cheminformatics

Springer

SpringerBriefs in Molecular Science

SpringerBriefs in Molecular Science present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields centered around chemistry. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic. Typical topics might include:

- A timely report of state-of-the-art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study
- A presentation of core concepts that students must understand in order to make independent contributions

Briefs allow authors to present their ideas and readers to absorb them with minimal time investment. Briefs will be published as part of Springer's eBook collection, with millions of users worldwide. In addition, Briefs will be available for individual print and electronic purchase. Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, easy-to-use manuscript preparation and formatting guidelines, and expedited production schedules. Both solicited and unsolicited manuscripts are considered for publication in this series.

Kunal Roy · Arkaprava Banerjee

q-RASAR

A Path to Predictive Cheminformatics

Springer

Kunal Roy
Drug Theoretics and Cheminformatics Laboratory
Department of Pharmaceutical Technology
Jadavpur University
Kolkata, West Bengal, India

Arkaprava Banerjee
Drug Theoretics and Cheminformatics Laboratory
Department of Pharmaceutical Technology
Jadavpur University
Kolkata, West Bengal, India

ISSN 2191-5407 ISSN 2191-5415 (electronic)
SpringerBriefs in Molecular Science
ISBN 978-3-031-52056-3 ISBN 978-3-031-52057-0 (eBook)
https://doi.org/10.1007/978-3-031-52057-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Foreword

Read-across is a field rapidly evolved in the last few years. Used for decades by the assessors to refine their judgment, it is widely applied for instance in the majority of the registration dossiers submitted to the European Authority for Chemicals (ECHA). For its wide use for authorization of substances, authorities such as ECHA (for the industrial chemical in Europe), EFSA (for food safety in Europe), SCCS (for cosmetic products in Europe), and US EPA (for chemicals in the USA) indagated the possible uses of read-across, introducing guideline documents, proposing specific tools, and promoting discussion about a formal way to proceed.

From a scientific perspective, read-across evolved from a manual exercise, done on a few substances, toward a more complex approach where multiple, heterogeneous kinds of data are used, and this requested the use of computers. The manual comparison of the structures has been substituted by the use of computer methods for similarity. This at the same time simplified the work and allowed for screening much larger databases. The structural similarity still remains the pillar of read-across, but other levels of information have been progressively added. Initially, the structural similarity was considered sufficient to perform the exercise, leaving to the expert the manual assessment associating other considerations. Today, the approaches, on a conceptual level and from a computational point of view as well, require further properties to be added and combined in the read-across exercise. The biological similarity, the toxicokinetic behavior, and the physico-chemical properties are some of the most popular ingredients in the modern read-across.

These multiple metrics for similarity increased the possibility of getting an accurate evaluation, exploring the different factors contributing to the behavior of the target substance. At the same time, this increased the level of complexity and opened new questions related to the way to integrate contributions deriving from eclectic components.

In some aspects, this kind of evolution is related to the evolution of the QSAR models. These models, initially very focused on specific groups of similar substances, evolved into tools with the ambition to cover all the substances, in principle. This required multiple descriptors, more data, and better algorithms.

The process of the evolution of the QSAR and read-across resulted in some cases of hybrid tools, merging the experience from the two areas, QSAR and read-across, originally separate.

The interest of the assessor is the evaluation of the substance. The technicalities associated with the specific approach, such as QSAR or read-across, may represent a barrier and not always a help for the assessor. It is time to move toward a closer integration between these separate tools and approaches, in the interest of exploiting at the best all data and methods and achieving a more accurate assessment of the substances.

Emilio Benfenati
Istituto di Ricerche Farmacologiche
Mario Negri IRCCS
Milan, Italy

Preface

Predictive cheminformatics is taking center stage in several applied fields including drug design, computational toxicology, materials science, agricultural science, nanosciences, food science, etc. Identifying the pattern governing the changes in responses (biological activity/toxicity/property) with changes in molecular structures and properties is the main focus of cheminformatics. Molecular similarity plays a very important role in understanding such relationships. Quantitative structure–activity/property relationships (QSARs/QSPRs) developed based on the molecular similarity principles have been used for predictive modeling, molecular design, and mechanistic interpretation of chemical–biological interactions for a long time. Read-across, a non-statistical data gap-filling approach used in chemical regulatory aspects, also uses the similarity concept not only concerning the chemical structures but also properties, toxicity, toxicokinetic, and toxicodynamics aspects. Although the original form of read-across is justification-based, recently several attempts have been made for the quantitative consideration of similarity and consequent predictions in a quantitative read-across approach. In the case of QSAR, the identification of important contributing features is an important step. Most of the transparent QSAR models are also able to quantify the contribution of individual descriptors toward the response being modeled. In the case of quantitative read-across, quantification of the contributions of individual features is not straightforward. On the other hand, quantitative read-across may also be applicable for the predictions from a limited number of source compounds, as it is a non-statistical approach. To combine the advantages of read-across and QSAR, recently these two techniques have been merged into a new field of quantitative read-across structure–activity relationship (q-RASAR). This approach uses various similarity- and error-based descriptors with or without the original chemical descriptors to generate QSAR-like models in a statistical framework. The similarity considerations are made in a machine learning approach while the final modeling approach may involve simple linear regression or the development of more sophisticated machine learning models. In the q-RASAR algorithm, RASAR descriptors are computed for a query compound not from its chemical structure or

property, but from its close congeners in the source set based on similarity considerations. Thus, the prediction aspect is introduced at the level of descriptor computation, and such models show superiority in the external prediction performance for various endpoints in comparison to the corresponding QSAR/QSPR models using the same level of chemical information. In addition, different RASAR descriptors are computed based on several chemical/biological descriptors, and thus they may work like latent variables of partial least squares regression problems. Hence, eventually, a lower number of RASAR descriptors may work well in comparison to a higher number of chemical regressors (chemical descriptors), especially when the training set size is limited. Due to an increase in the quality of external predictions, the q-RASAR modeling approach has been applied for several biological activities, toxicity, and materials property endpoints with demonstrated success. In addition to the regression problems, RASAR has also been applied for classification problems (classification RASAR or c-RASAR) and also for developing interspecies correlations (quantitative read-across structure-activity-activity relationships or RASAARs). The RASAR concept also addresses the aspects of activity cliffs and applicability domain through different similarity coefficients and plots. There is a provision for the application of q-RASAR in multi-endpoint and multi-target modeling and also to relate a particular biological effect caused by a chemical to a molecular initiating event in an adverse outcome pathway. Further exploration of the q-RASAR approach in modeling hitherto unexplored pharmaceutical endpoints is warranted. The progress in the q-RASAR research is being updated on the page https://sites.google.com/site/kunalroyindia/home/rasar. The present SpringerBrief will introduce 'q-RASAR' to the readers with a basic concept of cheminformatics and QSAR. We hope that the researchers in the field will find this novel approach interesting for further modeling of various complex biological or non-biological endpoints.

Kolkata, India
November 2023

Kunal Roy
Arkaprava Banerjee

Acknowledgments The q-RASAR research is being carried out at the Drug Theoretics and Cheminformatics (DTC) Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India, with funding from Life Sciences Research Board (LSRB), Defence Research and Development Organisation (DRDO), Government of India (LSRB/01/15001/M/LSRB-394/SH&DD/2022).